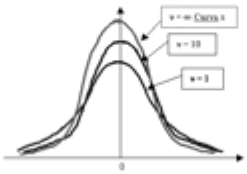
Probabilidad y Estadística 9

Actividades de Aprendizaje



Conceptos y definiciones de esta clase:

- Inferencia estadística
- Diferencia entre parámetros y estadísticos
- La media aritmética como estimador
- Propiedades de un buen estimador
- Distribuciones en el muestreo
- El teorema del límite central
- Estimación Puntual y por Intervalo de Confianza
- Cálculo del intervalo de confianza para la media, conocida la desviación típica de la población en una variable aleatoria normal

1. Inferencia estadística

Hasta aquí nos hemos iniciado en el arte de recopilar datos, agruparlos, ordenarlos y sacar algunas conclusiones sobre ellos. Por otro lado, hemos aprendido a calcular las probabilidades que tiene un evento de comportarse de una determinada manera. Y también nos hemos aplicado a la tarea de analizar las propiedades y los usos de los diversos estimadores.

La pregunta natural que surge es: ¿podríamos utilizar estas u otras herramientas similares para predecir lo que pasará en un conjunto de elementos a partir de una muestra que resulte representativa del todo? Las razones que podríamos tener para necesitarlas son muy variadas. Por ejemplo, si vamos a lo estrictamente económico, a menudo cuesta mucho dinero realizar encuestas y procesar los datos obtenidos. Otras veces la medición de la totalidad de los elementos resulta imposible de practicar porque estos se generan constantemente, tal como ocurre en una línea de producción. También ocurre, en algunos casos, que la evaluación que se está llevando a cabo implica la modificación o incluso destrucción del elemento observado, como ocurre en las pruebas industriales de resistencia.

1.1 ¿Qué significa inferir?

Inferir algo es sacar conclusiones a partir de una muestra. Y de eso se trata esta unidad: de asegurarnos que estamos tomando una muestra adecuada y representativa de un todo que deseamos analizar para luego estudiar las mejores y más adecuadas estrategias para sacar conclusiones de antemano.

Este tipo de herramientas, junto con las de predicción que serán vistas más adelante, son muy útiles en la medicina, la economía, las ciencias sociales, la meteorología, el campo de las actividades políticas y militares, la astronomía, la biología y casi todas

las actividades industriales. La ventaja y el poder de estar en condiciones de sacar conclusiones de antemano son ilimitados, y permiten ahorrar costos en investigación y desarrollo, ponderar la influencia de las modificaciones a incorporar en un proceso ya analizado, o incluso aventurarse a definir y poner a prueba nuevas estrategias sin que ello derive necesariamente en altos costos de modelización, producción y puesta en marcha de proyectos.

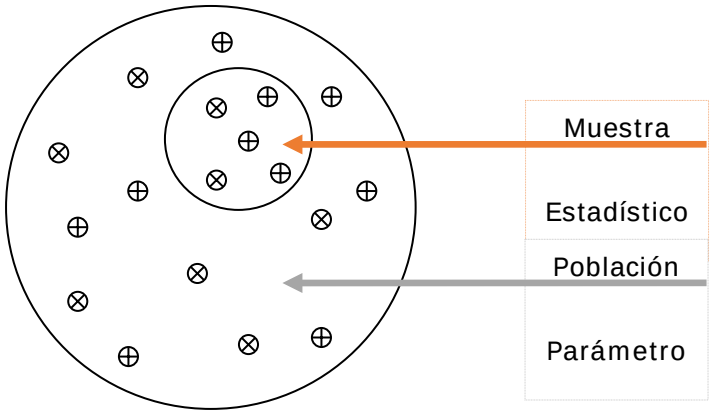
1.2 Diferencia entre parámetros y estadísticos

Básicamente diremos que, a diferencia de lo estudiado anteriormente, no realizaremos cálculos sobre la totalidad de los elementos de una población (análisis mediante **parámetros**), sino más bien sobre una muestra confiable y representativa del total (análisis mediante **estadísticos**).

Estos estadísticos muestrales (o estadísticos, a secas), son cuantificadores de los datos de una muestra, que se calculan con el objetivo de inferir conclusiones sobre la población de la que ha sido extraída la mencionada muestra. Por ejemplo, ´los estimadores $\bar{X}$ (media muestral) y S^2 (varianza muestral) son los estadísticos correspondientes a los parámetros μ (media poblacional) y σ^2 (varianza), respectivamente.

Tenemos numerosos ejemplos de esto en nuestra vida cotidiana. ¿Quién no se ha realizado un análisis de sangre? Pues bien, a nadie se le ocurriría que, para realizar un conteo de glóbulos blancos, deberíamos extraer toda la sangre de un individuo. De esto se trata todo lo que estudiaremos a continuación, y que consiste en el objetivo principal de la **Estadística Inferencial**: *la generalización de conclusiones sobre una población, a partir del estudio de los datos obtenidos en una muestra*.

Podríamos visualizarlo de la siguiente forma:



El estadístico calculado de la muestra sirve para estimar el parámetro poblacional

Ejemplo 1

Analicemos algo muy habitual en la actividad de desarrollo de sistemas. Un analista programador recibe los lineamientos para confeccionar una serie de listados sobre datos de un sistema, y debe entregar un presupuesto de tiempo y costos que demandará lo solicitado. Además, como siempre ocurre en sistemas, debe entregar su presupuesta para esa misma tarde.

El analista se encuentra frente a un problema de inferencia: debe calcular cuánto tiempo demandará la tarea basado en su experiencia previa, o consultando a sus pares. Luego, deberá valuar la tarea, basado en el precio promedio de la hora de programación.

1.3 La media aritmética como estimador

Un buen estimador que puede utilizar es la media aritmética, que cuando la aplicamos a una muestra, se denomina **media muestral** y es uno de los **estadísticos muestrales**. Se trata de un estimador fiable, fácil de calcular y conocido por todos.

Entonces, el programador toma algunos de los listados asignados (una muestra) y consulta a sus compañeros cuánto suponen que les demandará la tarea. Luego, calcula el promedio de los datos obtenidos, y a partir del mismo, infiere la cantidad que le demandará el total. Por otro lado, realiza el mismo procedimiento con el costo de la hora de programación y finalmente multiplica ambos resultados.

En fórmula, su cálculo será

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

En donde $\bar{X}$ es la media muestral que se desea obtener, n es la cantidad de elementos contenidos en la muestra y los x_i son cada uno de los valores de la muestra.

Casi de inmediato nos surgen dudas y notamos las debilidades que puede tener este procedimiento. Puede pasar, por ejemplo, que el analista haya seleccionado justo a los programadores más experimentados, o a los más caros, y de esta manera su estimación pudo desvirtuarse. Es por eso que, para minimizar el impacto de ello, se analizan las propiedades deseables para que un estimador sea considerado como adecuado.

Denotaremos con $\hat{\theta}$ al estimador genérico de un determinado parámetro θ. Cuando se repiten las muestras para calcular el estimador, a cada uno de ellos se lo suele denominar $\hat{\theta}_n$, en donde n a menudo puede indicar la cantidad de elementos considerados en la muestra.

1.4 Propiedades de un buen estimador

1. ***Un buen estimador es insesgado***

El sesgo de un estimador es la diferencia que hay entre la esperanza matemática del estimador y el valor hallado para el mismo. Se considera que un buen estimador tiene sesgo nulo o casi nulo. Recordemos que la esperanza es el valor esperado para una determinada prueba. Simbólicamente resulta:

$$E(\hat{\theta}) = \theta$$

Ejemplo 2

Si se usa la media muestral $\bar{X}$ para estimar la media poblacional μ de una población con distribución normal, se sabe que la $\mu_x = \mu$, por lo tanto, la media es un estimador insesgado.

Cabe aclarar que a menudo no se espera que $\hat{\theta}$ coincida exactamente con θ, sino que se espera que no esté muy alejado de su valor. Si el estimador no es insesgado, entonces a la diferencia $\theta - E(\hat{\theta})$ se la conoce como sesgo del estimador $\hat{\theta}$.

2. *Un buen estimador es consistente*

Se dice que un estimador es consistente cuando, a medida que crece el tamaño de la muestra, su valor se aproxima cada vez más al valor del parámetro considerado. Simbólicamente:

$$\lim_{n\to\infty} \hat{\theta} = \theta$$

Si $\hat{\theta}$ es un estimador insesgado de θ, y la varianza del estimador tiende a cero bajo las mismas condiciones, se deduce entonces que el estimador es consistente, vale decir:

$$\lim_{n\to\infty} \sigma^2(\hat{\theta}) = 0$$

Ejemplo: la media muestral $\bar{X}$ es un estimador consistente de la media poblacional en una distribución normal, ya que la varianza de la misma tiende a cero cuando $n \to \infty$, lo cual se manifiesta gráficamente en que la distribución se concentra alrededor del verdadero valor □ a medida que n crece.

3. *Un buen estimador es eficiente*

Un estimador es más eficiente, o más preciso, cuanto menor sea su varianza. Usualmente se utiliza esta característica en la comparación de dos o más estimadores, o sea:

$$\sigma^2(\theta_1) < \sigma^2(\theta_2) \Rightarrow \theta_1 \; es \; más \; eficiente \; que \; \theta_2$$

4. Un buen estimador es suficiente

Un estimador es suficiente si para calcularlo se utiliza toda la información relevante suministrada por la muestra.

1.5 Distribuciones en el muestreo

Una estadística es una variable aleatoria que depende básicamente de la muestra seleccionada, con lo cual debe tener una distribución de probabilidad, a la que llamaremos **distribución muestral**, la cual, a su vez, dependerá de:

- *El tamaño de la población*
- *El tamaño de la muestra*
- *El método utilizado para la selección de las muestras*

Un caso concreto con el que estamos familiarizados – *y conocemos a menudo las limitaciones y hasta los desaciertos que brinda* – son las encuestas electorales. Y en contraposición a las mismas, siempre sobre el mismo tema, conocemos las encuestas elaboradas a "boca de urna", las cuales habitualmente suponen un margen muy bajo de error. Justamente, algunos de los factores determinantes en la certeza de estos métodos, tiene que ver, precisamente, con lo que analizaremos a continuación.

Distribución en el muestreo de la media

Como su nombre lo indica, es la distribución muestral correspondiente a las medias de las diversas muestras seleccionadas. Si, por ejemplo, tomamos una muestra aleatoria de tamaño n de una población de distribución normal con media μ y varianza σ^2, entonces cada observación de esta muestra $(x_1, x_2, \ldots, x_n)$ es una variable aleatoria distribuida normal e independientemente, cuya media será μ y su varianza σ^2.

De esta manera podemos decir que la media muestral $\bar{X}$ tiene una distribución normal con media μ y su varianza σ^2/n .

Lo llamativo es que, si se muestrea una población que tiene una distribución de probabilidad desconocida, la distribución de muestreo de la media muestral seguirá siendo aproximadamente normal con media μ y varianza σ^2/n , si el tamaño de la muestra n es grande. Esto es lo que se conoce como **teorema del límite central**, y que se enuncia a continuación.

1.6 El teorema del límite central

Sea $(x_1, x_2, \ldots, x_n)$ una muestra de n variables aleatorias, independientes e idénticamente distribuidas de una distribución con media μ y varianza finita σ^2 finita y distinta de cero, y donde $\bar{X}$ es la media muestral, entonces la forma límite de la distribución de

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

cuando $n \to \infty$, es la distribución normal estándar.

1.7 Estimación Puntual y por Intervalo de Confianza

La estimación puntual es la estimación del valor de un parámetro mediante un único valor, obtenido mediante una fórmula determinada. Por ejemplo, si se pretende estimar el peso medio de un determinado grupo, y entonces se extrae una muestra y se calcula el peso medio de esa muestro, para finalmente inferir el peso medio de la población en base a esta medida.

La media muestral es un es un estimador puntual de la media poblacional.

Como vimos en puntos anteriores, no podemos esperar que una estimación puntual nos brinde exactamente el valor correcto de la población. Es más, si tomamos varias muestras para realizar la estimación correspondiente, vamos a obtener valores distintos. Para subsanar este problema, se suele sumar y/o restar una cantidad (llamada margen de error) al estimador puntual, para así calcular una estimación por intervalos.

Estimación por intervalo = Estimación Puntual $\pm$ Margen de Error

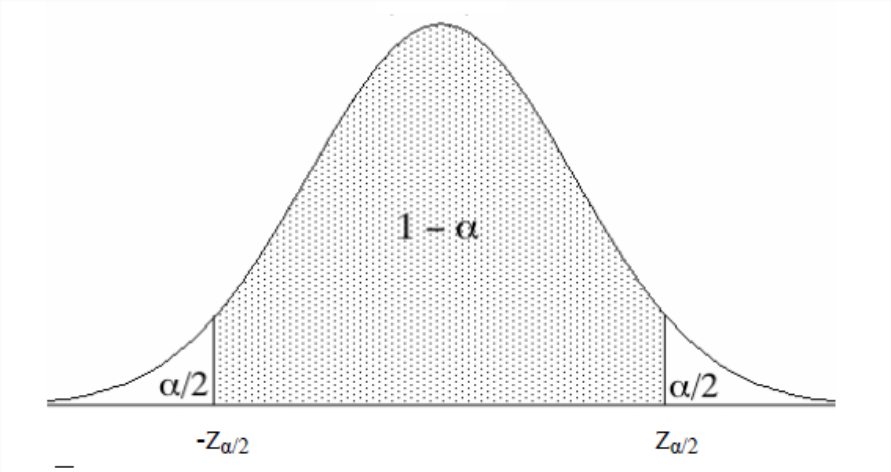
Estos intervalos estarán relacionados con el tipo de distribución que presente la media. Así, por ejemplo, si la media $\bar{x}$ presenta una distribución normal, el 95% de los valores de $\bar{x}$ se encontrarán dentro de $\pm 1.96\,\sigma$ de la media poblacional μ, tal como se explica a continuación.

1.8 Cálculo del intervalo de confianza para la media, conocida la desviación típica de la población en una variable aleatoria normal

Sabiendo que la media muestral $\bar{x}$, sigue una distribución normal de media μ y desviación típica $\sigma/\sqrt{n}$, se calculan los extremos del intervalo como:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \quad \text{y} \quad \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

Gráficamente, el área sombreada nos indica el intervalo buscado:



Actividad para la Facilitación de los Aprendizajes

Situación Problemática 1

Se realizan 10 tomas de los tiempos utilizados por un corredor para realizar una prueba de velocidad, con el propósito de estimar la media del tiempo empleado por el atleta.

El resultado de las medidas tomadas arroja una media de 21,5 segundos.

Sabiendo por datos anteriores que la desviación típica de esta variable para este corredor es de 0,4 segundos, obtener un intervalo de confianza con un 95% de confianza. ¿Cuántas pruebas habría que cronometrar para que el margen de error en la estimación de la media fuese inferior a tres décimas de segundo?

NOTA: Se establece que la variable que mide el tiempo del corredor sigue una distribución normal.

Solución:

Los datos con los que contamos son:

$\bar{x} = 21.5$ seg.

$n = 10$

$\sigma = 0.4$ seg.

$1 - \alpha = 0.95$

Buscaremos entonces el valor de $z_{\alpha/2}$ para una distribución normal que tenga un área de probabilidad igual a $\alpha/2$ por su derecha, que en nuestro caso equivale a

(1-0.95) /2, es decir 0.025.

La función de distribución de probabilidad nos da el área de probabilidad acumulada, es decir a la izquierda, con lo cual tenemos que ver qué valor de z nos deja a la izquierda 0,975, y este valor se corresponde con un z=1.96.

Con estos datos podremos calculas los extremos del intervalo:

$$\left(21.5 - \frac{0.4}{\sqrt{10}}1.96\,, 21.5 + \frac{0.4}{\sqrt{10}}1.96\right)$$

$$(21.5 - 0.248\,, 21.5 + 0.248)$$

$$(21.252\,, 21.748)$$

Con respecto a la pregunta que se realiza, el cálculo sería:

$$\frac{\sigma}{\sqrt{n}} z_{\alpha/2} < 0.3$$

$$\frac{0.4}{\sqrt{n}} 1.96 < 0.3$$

$$n > 6.82$$

Interpretación de los resultados:

Se estima que la media es de 21.5 seg., más o menos un error de 0.248 seg.